A Survey of Association Rule Based Techniques for Preserving Privacy and Security

Rajnik S Katriya[1], Neha Gupta[2] , Ashish Patel[3]
[1,2]*Computer Engineering Department, Silver Oak College of Engg and Technology, Ahmedbad*
[3]*Computer Engineering Department, Charotar University of Science and Technology, Changa*

Abstract- Data mining is that the extraction of attention-grabbing patterns or data from large quantity of information. In recent years, with the explosive development in web, knowledge storage and processing technologies, privacy preservation has been one in every of the larger issues in data processing. variety of ways and techniques are developed for privacy conserving data processing. This paper provides a good survey of various privacy conserving data processing algorithms and analyses the representative techniques for privacy conserving data processing, and points out their deserves and demerits. Finally the current issues and directions for future analysis area unit mentioned.

Index Terms- Privacy Preserving Data Mining, Privacy, Randomization, K-Anonymity, Secure Multiparty Computation

I. INTRODUCTION

Data mining may be a well-known technique for mechanically and showing intelligence extracting data or knowledge from an oversized quantity information, however, it can even revelation sensitive data regarding individuals compromising the individual’s right to privacy [1]. A number of effective strategies for privacy preserving data processing are planned [2]. The problem of privacy conserving data processing and security has become a lot of necessary in recent years attributable to the increasing ability to store personal knowledge concerning users and therefore the increasing sophistication of mining algorithmic rule to leverage this information. variety of techniques like classification, k- obscurity, association rule mining, cluster are steered in recent years so as to perform privacy conserving data processing. moreover, the matter has been mentioned in multiple communities like the information community, the applied math revealing management community and therefore the cryptography community. data processing techniques are developed with success to extracts data so as to support a spread of domains promoting, foretelling,medical designation, and national security. however it's still a challenge to mine sure varieties of knowledge while not violating the information house owners privacy .

This paper provides a large survey of various privacy protective data processing techniques and analyses the representative strategies for privacy protective data processing, and points out their deserves and demerits. the remainder of this paper is organized as follows. In section 2, we are going to introduce the Classification of Techniques for safeguarding Sensitive knowledge. In section 3, we are going to analyze Techniques of privacy protective. In section 4, we are going to analyze the strategy of randomization for privacy protective on the initial knowledge. In section five, we are going to discuss the anonymization technique. The encryption technique are mentioned in section 6. In section 7, we are going to discuss the analysis of Techniques for safeguarding Sensitive information. Section 8 contains the conclusions and future work.

II. CLASSIFICATION OF TECHNIQUES FOR PROTECTING SENSITIVE DATA

There square measure several approaches that are adopted for privacy protective data processing. we are able to classify them supported the subsequent dimensions:

- knowledge distribution
- knowledge modification
- data processing rule
- knowledge or rule concealing
- privacy preservation

The knowledge distribution refers to the distribution of data. a number of the approaches are developed for centralized knowledge, whereas others visit a distributed knowledge situation. Distributed knowledge situations also can be classified as horizontal knowledge distribution and vertical knowledge distribution. Horizontal distribution refers to those cases wherever completely different information records reside in several places, whereas vertical knowledge distribution, refers to the cases wherever all the values for various attributes reside in several places.

The knowledge modification refers to the information modification theme. In general, knowledge modification is employed so as to change the first values of a information that has to be free to the general public and during this thanks to guarantee high privacy protection. it's necessary that an information modification technique ought to be together with the privacy policy adopted by a company. Methods of modification include:

- Perturbation, that is accomplished by the alteration of an attribute worth by a brand new worth (i.e., dynamical a 1-value to a 0-value, or adding noise),
- Blocking, that is that the replacement of an existing attribute worth with a “?”,
- Aggregation or merging that is that the combination of many values into a coarser class,
- Swapping that refers to interchanging values of individual records, and
- Sampling, this refers to emotional information for less than a sample of a population. The dimension refers to the info mining algorithmic program, that the information modification is happening. this can be really one thing that's not familiar beforehand, however it facilitates the analysis and style of the info activity algorithmic program. For the nowadays, numerous data processing algorithms are thought of in isolation of every different. Among them, the foremost necessary concepts are developed for classification data processing algorithms, like call tree inducers, association rule mining algorithms, agglomeration algorithms, rough sets and Bayesian networks.

The knowledge or rule concealing refers as to if information or aggregate data ought to be hidden. The quality for activity aggregate information within the style of rules is after all higher, and for this reason, largely heuristics are developed. The change of the number of public data causes the information mineworker to provide weaker logical thinking rules that may not enable the logical thinking of confidential values. This method is additionally called “rule confusion”.

The privacy preservation that is that the most significant refers to the privacy preservation technique used for the selective modification of the information. Selective modification is needed so as to realize higher utility for the changed information provided that the privacy isn't jeopardized.

III. TECHNIQUES OF PRIVACY PRESERVING

The goal of privacy conserving data processing is to develop data processing ways while not increasing the danger of misuse of the info accustomed generate those ways. the subject of privacy conserving data processing has been extensively studied by the info mining community in recent years. variety of effective ways for privacy conserving data processing are planned. Most ways use some style of transformation on the initial information so as to perform the privacy preservation. The reworked dataset is created obtainable for mining and should meet privacy necessities while not losing the advantage of mining. we tend to classify them into the subsequent 3 categories:

The Randomization method. **(Reconstruction-Based Techniques)**

Randomization method could be a widespread methodology in current privacy conserving data processing studies. It masks the values of the records by adding noise to the first information.The noise side is sufficiently massive so the individual values of the records will not be recovered. However, the likelihood distribution of the mixture information are often recovered and later on used for privacy-preservation functions. In general, randomization methods aims at finding associate degree acceptable balance between privacy preservation and information discovery. Representative randomization methods ways embody random-noise-based perturbation and irregular Response theme.

Randomization method is additional economical. However, it leads to high data loss.

The Anonymization method

Anonymization method aims at creating the individual record be indistinguishable among a bunch records by mistreatment techniques of generalization and suppression. The representative Anonymization method is k-anonymity. The motivating issue behind the k-anonymity approach is that several attributes within the information will typically be thought of quasi-identifiers which might be employed in conjunction with public records so as to unambiguously establish the records. several advanced strategies are planned, such as, p-sensitive k-anonymity, (a, k)-anonymity, l-diversity, t-closeness, M-invariance, personalised obscurity, and so on. The Anonymization method will make sure that the reworked information is true, however it conjointly leads to data loss in some extent.

The Encryption method. **(Cryptography-Based Techniques)**

Encryption method in the main resolves the issues that individuals collectively conduct mining tasks supported the personal inputs they supply. These mining tasks might occur between mutual un-trusted parties, or maybe between competitors, therefore, protective privacy becomes a primary concern in distributed data processing setting. There square measure 2 totally different distributed privacy conserving data processing approaches like the strategy on horizontally divided information which on vertically divided information. The Encryption method will make sure that the reworked information is actual and secure, however it's a lot of low economical.

IV. THE RANDOMIZATION METHOD (RECONSTRUCTION-BASED TECHNIQUES)

The randomization method provides a good however easy means of preventing the user from learning sensitive knowledge, which might be simply enforced at knowledge assortment part for privacy protective data processing, as a result of the noise additional to a given record is freelance of the behavior of alternative knowledge records. once the randomization method is distributed, the info assortment method consists of 2 steps [3]. the primary step is for the knowledge suppliers to disarrange their knowledge and transmit the randomized data to the information receiver. within the second step, the info receiver estimates the first distribution of the info by using a distribution reconstruction algorithmic rule. The model of randomization is shown in Figure 1.

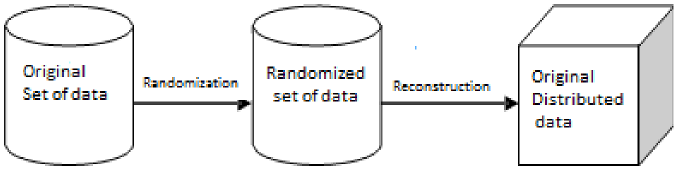

Figure 1. The Randomization Model

Representative randomization ways embrace random-noise-based perturbation and irregular Response theme. Agrawal and Srikant projected a theme for privacy protective data processing victimization random perturbation and mentioned however the reconstructed distributions could also be used for data processing [4]. In their randomization scheme, a random variety is further to the worth of a sensitive attribute. as an example, if xi is that the worth of a sensitive attribute, **xi + ri** instead of **xi** can seem within the information, wherever **ri** may be a random noise drawn from some distribution. it's shown that given the distribution of random noises, reconstructing the distribution of the initial information is feasible. later, Evmievski et al. planned associate approach to conduct privacy protective association rule mining [5]. Kargupta et al. [6] planned a random matrix-based spectral filtering technique to recover the initial information from the discomposed information. Huang et al. additional projected 2 alternative information reconstruction methods: PCA-DR and MLE-DR in [7]. Guo et al. self-addressed the difficulty of providing accuracy in terms of varied reconstructed measures in privacy protective market basket information analysis [8]. The randomization method may be a straightforward technique which may be simply enforced at information assortment time. it's been shown to be a helpful technique for concealment individual information in privacy protective data processing.

The method methodology is additional economical. However, it leads to high data loss.

V. THE ANONYMIZATION METHOD

With the ascension in information, networking, and computing technologies, an outsized quantity of private knowledge are often integrated and analyzed digitally, resulting in associate degree inflated use of information mining tools to infer trends and patterns. This has raised universal considerations regarding protective the privacy of people. the information records square measure usually created out there by merely removing key identifiers like the name and social-security numbers from individual records. However, the combos of different record attributes are often wont to specifically determine individual records. for instance, attributes like race, birth, sex, and nothing square measure accessible publicly records like citizen list. once these attributes also are accessible during a given knowledge set like medical knowledge, they will be wont to infer the identity of the corresponding individual with high likelihood by linking operation.

In recent years, various algorithms are planned for implementing k-anonymity via generalization and suppression. Bayardo associate degreed Agrawal [9] given an optimum algorithmic rule that starts from a completely generalized table and specializes the dataset during a bottom k-anonymous table. LeFevre et al. [10] represented associate degree algorithmic rule that uses a bottom-up technique and a priori computation. Fung et al. [11] given a top-down heuristic to form a table to be free k-anonymous. on the theoretical results, Sweeney [12] proved the optimum k-anonymity is NP-hard and provided approximation algorithms for optimum k-anonymity. However, Machanavajjhala et al. [13] noticed that the user might guess the sensitive values with high confidence once the sensitive knowledge is lack of diversity.

The k-anonymity ways in the main concentrate on a universal approach that exerts an equivalent quantity of preservation for all people, while not business for his or her concrete wants. The consequence is also providing insufficient protection to a set of individuals, whereas applying excessive privacy management to a different set. actuated by this, Xiao and Tao [14] given a brand new generalization framework supported the conception of customized obscurity. Their technique performs the minimum generalization for satisfying everybody's needs, and thus, retains the most important quantity of information from the first data. additionally, the present k-anonymity solutions supported generalization and suppression techniques suffer from high info loss and low usability principally because of reliance on pre-defined generalization hierarchies or full order obligatory on every attribute domain.

k-Anonymity data processing is but a recent analysis space and lots of problems are still to be investigated, such as, the mixture of k-anonymity with different potential data processing techniques; the investigation of latest approaches for police investigation and block k-anonymity violations. The anonymization methodology will make sure that the reworked knowledge is true, however it additionally leads to info loss in some extent.

VI. THE ENCRYPTION METHOD FOR PRIVACY PRESERVING IN DISTRIBUTED ENVIRONMENT (CRYPTOGRAPHY-BASED TECHNIQUES)

The growth of web has triggered tremendous opportunities for distributed data processing, wherever individuals collectively conducting mining tasks supported the non-public inputs they provides. These mining tasks may occur between mutual un-trusted parties, or maybe between competitors, therefore, protective privacy becomes a primary concern in distributed data processing setting. Distributed privacy protective data processing algorithms need collaboration between parties to reckon the results or share no-sensitive mining results, whereas demonstrably resulting in the revelation of any sensitive data.

In general, distributed data processing involves two forms: horizontally partitioned data and vertically partitioned data. Horizontally partitioned data means every web site has complete data on a definite set of entities, and an integrated dataset consists of the union of those datasets. In distinction, vertically partitioned data has differing kinds of data at every site; every has partial information on constant set of entities. Most privacy protective distributed data processing algorithms area unit developed to reveal nothing apart from the ultimate result. Kantarcioglu and Clifton [15] studied the privacy-preserving association rule mining drawback over horizontally partitioned data. Their strategies incorporate cryptologic techniques to reduce the data shared, whereas adding very little overhead to the mining task. Lindell et al. the matter of in private mining association rules on vertically partitioned data was addressed in [16,17]. Vaidya and Clifton initial studied however secure association rule mining is finished vertically partitioned data by extending the Apriori algorithmic rule. these strategies area unit nearly supported the special secret writing protocol called Secure Multiparty Computation (SMC) technology. [18] provide a secure algorithmic rule of the distributed association rule mining, which might stop effectively the collusion of parties by using the Shamir’s secret sharing technique, and provide the algorithmic rule with respect to its security, efficiency and correctness. The SMC literature defines 2 basic adversarial models:

Semi-Honest Model: Semi-honest adversaries follow the protocol dependably, however will attempt to infer the key info of the opposite parties from the info they see throughout the execution of the protocol.

Malicious Model: Malicious adversaries could do something to infer secret data. they'll abort the protocol at any time, send spurious messages, spoof messages, interact with alternative (malicious) parties, etc. SMC technology employed in distributed privacy conserving data processing areas in the main consists of a group of secure sub-protocols, such as, secure sum, secure comparison, scalar product protocol, secure intersection, secure set union utilized in on. within the following, we'll in short describe the essential plan of 2 forms of secure sub-protocols employed in horizontally partitioned and vertically partitioned setting.

Secure Sum Protocol: Secure total will firmly calculate the add of values from completely different sites. Assume that every website I has some value vi and every one sites wish to firmly calculate S=v1+v2+v3+…..+vn, wherever is thought to be within the vary [0..m]. for instance, in horizontally partitioned association rule mining setting, we are able to firmly calculate the worldwide support count of an itemset by the secure add sub-protocol.

Dot Product Protocol: To present, several secure scalar product protocols are planned. the matter will be outlined as follows: Alice features a n-dimensional vector X=(x1,x2,x3..xn), whereas Bob incorporates a n-dimensional vector Y=(y1,y2,y3…yn). At the top of the protocol, Alice ought to get ra=(X*Y+rb) wherever range may be a random number chosen from uniform distribution that's acknowledged solely to Bob, and X*Y=(x1y1+x2y2+x3y3+….+xnyn) . for instance, exploitation the scalar product protocol square measure able to} firmly calculate the worldwide support count of an itemset whose items are set at totally different sites in vertically data setting.

The secret writing technique will make sure that the reworked knowledge is actual and secure, however it's a lot of low economical. Moreover, most existing work on terribly efficient privacy protective data mining solely provides the protocols against semi-honest adversaries. a very important space for future analysis is to develop economical mining protocols that stay secure and personal even though a number of the parties concerned behave maliciously.

VII. EVALUATION OF TECHNIQUES FOR PROTECTING SENSITIVE DATA

An important side within the development and assessment of algorithms and tools, for privacy conserving data processing is that the identification of appropriate analysis criteria and also the development of connected benchmarks. it's typically the case that no privacy conserving algorithmic rule exists that outperforms all the others on all potential criteria. Rather, an algorithmic rule might perform higher that another one on specific criteria, like performance and/or knowledge utility. it's therefore necessary to produce users with a group of metrics which is able to modify them to pick the foremost acceptable privacy conserving technique for the information at hand; with relevance some specific parameters they're inquisitive about optimizing.

A preliminary list of analysis parameters to be used for assessing the standard of privacy conserving data processing algorithms, is given below:

• The performance of the planned algorithms in terms of your time necessities, that's the time required by every algorithmic rule to cover a such as set of sensitive information;

• the information utility when the appliance of the privacy conserving technique, that is equivalent with the minimization of the knowledge loss as an alternative the loss within the practicality of the data;
• the extent of uncertainty with that the sensitive data that are hidden will still be expected

VIII. CONCLUSIONS

With the event of information analysis and process technique, the privacy revealing drawback concerning individual or company is inevitably exposed once emotional or sharing information to mine helpful call data and information, then provide the birth to the analysis field on privacy protective data processing. during this paper,we given totally different problems and repeat naïve privacy conserving strategies to distribute ones and also the strategies for handling horizontally and vertically partitioned off information. whereas all the purposed strategies square measure solely approximate to our goal of privacy preservation, we'd like to additional good those approaches or develop some economical strategies. To address these problems, following drawback ought to be wide studied.

1. In distributed privacy conserving data processing areas, potency is a necessary issue. we must always attempt to develop a lot of efficient algorithms and attain a balance between revealing value, computation value and communication cost.
2. Privacy and accuracy could be a try of contradiction; rising one sometimes incurs a price within the different. a way to apply varied optimizations to realize a trade-off ought to be deeply researched.
3. Side-effects area unit inevitable in information cleansing method. a way to scale back their negative impact on privacy protective must be thought-about rigorously. we tend to additionally ought to outline some metrics for measuring the side-effects resulted from processing.

REFERENCES

[1] Han Jiawei, M. Kamber, Data Mining: Concepts and Techniques, Beijing: China Machine Press, pp.1-40,2006.

[2] V.S.Verykios, E.Bertino, I.N.Fovino, L.P.Provenza, Y.Saygin, Y.Theodoridis, “State-of-the-art in Privacy Preserving Data Mining”, New York, ACM SIGMOD Record, vol.33, no.2, pp.50-57,2004.

[3] N. Zhang, “Privacy-Preserving Data Mining”, Texas A&M University, pp.19-25, 2006.

[4] R. Agrawal, R. Srikant, “Privacy-Preserving Data Mining”, ACM SIGMOD Record, New York, vol.29, no.2, pp.439-450,2000.

[5] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, “Privacy Preserving Mining of Association Rules”, Information System, vol.29, no.4, pp.343-364,2004.

[6] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, “On the Privacy Preserving Properties of Random Data Perturbation Techniques”, In Proceedings of the 3rd International Conference on Data Mining, pp.99-106, 2003.

[7] Z. Huang, W. Du, B. Chen, “Deriving Private Information from Randomized Data”, In Proceedings of the ACM SIGMOD Conference on Management of Data, Baltimore, Maryland, USA, pp.37-48, 2005.

[8] L. Guo, S. Guo, X. Wu, “Privacy Preserving Market Basket Data Analysis”, In Proceedings the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, pp.103-114, 2007.

[9] R. Bayardo, R. Agrawal, “Data Privacy Through Optimal k-Anonymization”, In Proceedings the 21st International Conference on Data Engineering, pp.217-228, 2005.

[10] K. Lefevre, J. Dewittd, R. Ramakrishnan, “Incognito: Efficient Full-Domain k-Anonymity”, In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp.49-60, 2005.

[11] B. Fung, K. Wang, P. Yu, “Top-down Specialization for Information and Privacy Preservation”, In Proceedings of the 21st IEEE International Conference on Data Engineering, pp.205-216, 2005.

[12] L. Sweeney, “Achieving k-Anonymity Privacy Protection Using Generalization and Suppression”, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol.10, no.5,pp.571-588,2002.

[13] A. Machanavajjhala, J. Gehrke, D. Kifer, “l-Diversity: Privacy Beyond k-Anonymity”, ACM Transactions on Knowledge Discovery from Data, pp.24-35,2007.

[14] X.K. Xiao, Y.F. Tao, "Personalized Privacy Preservation", In Proceedings of the ACM Conference on Management of Data (SIGMOD), pp.229-240, 2006.

[15] M. Kantarcioglu, C. Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data", IEEE Transactions on Knowledge and Data Engineering, vol.16, no.9, pp.1026-1037, 2004.

[16] J. Vaidya, C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data", In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.639-644, 2002.

[17] J. Vaidya, C. Clifton, "Privacy-Preserving k-Means Clustering over Vertically Partitioned Data", In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.206–215, 2003.

[18] Xinjing Ge; Li Yan; Jianming Zhu; Wenjie Shi, "Privacy-preserving distributed association rule mining based on the secret sharing technique," Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on , vol., no., pp.345,350, 23-25 June 2010